ENVRI Common Operations of Environmental Research Infrastructures

OEILM: an ontological framework for environmental big data research infrastructures

Zhiming Zhao
z.zhao@uva.nl
System and Network Engineering,
University of Amsterdam (UvA)

Project number: 283465

Environmental problems

Global warming

Earthquakes

Climate change

Fresh water

Deforestification

Volcanoes

Biodiversity loss

Epidemic diseases

Food supplies

Pollution

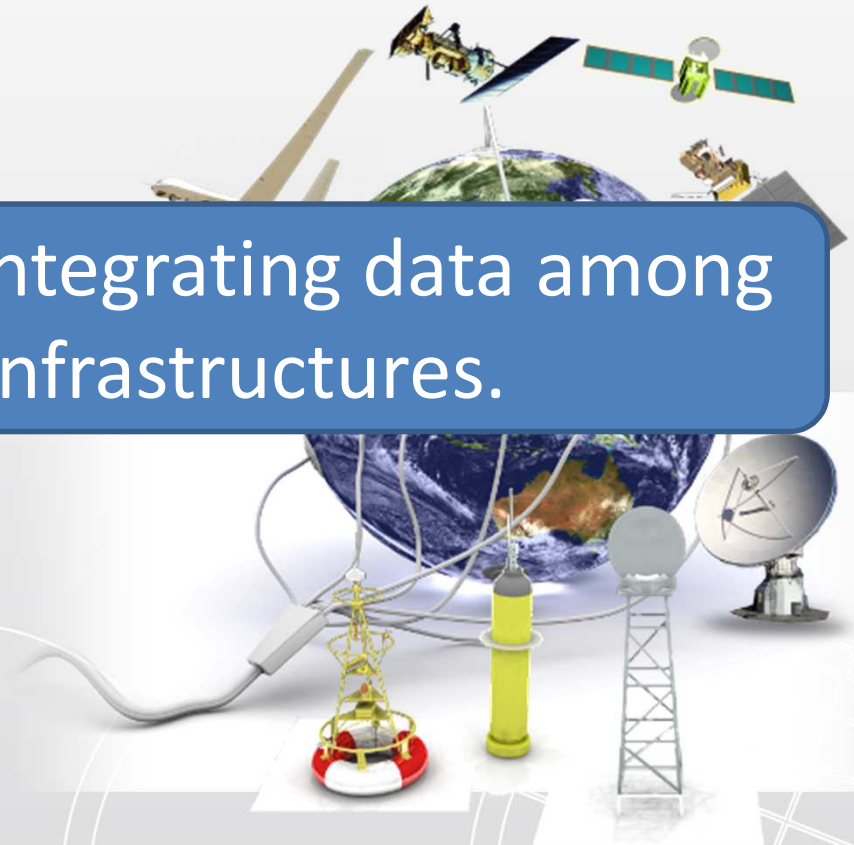
Environmental data and research infrastructure

- **Environmental data:**
  - Observation and measurement, time and locations
  - carbon
  - plate
  - seafloor
  - Space weather,
  - ocean
  - bio diversity
  - …

Environmental data and research infrastructure

- **Environmental data:**
  - Observation and measurement, time and locations
  - carbon
  - plate
  - seafloor
  - …

Difficulties in sharing and integrating data among different research infrastructures.



Outline

- Environmental research infrastructures
- ENVRI research project
- Open Environmental Information Linking Model (OEILM)
- Use cases
- Discussions

Environmental research infrastructure

- **An environment research infrastructure**
  - **Acquisition --** brings the measures/data streams into the system (**non-reproducible** data)
  - **Curation --** manages/maintains quality data (**reproducible** data)
  - **Access --** facilities discovery, access (**published** data)
  - **Processing --** facilities analysis/mining/experiments (**combined/derived** data)
  - **Community Support --** supports users to conduct their roles in communities (**user generated** data)
- **ESFRI: European strategic forum of research infrastructure**

EISCAT-3D	EMSO	EPOS	EURO-ARGO	ICOS	LIFEWATCH
• Upgrade of incoherent SCATter facility	• Multidisciplinary seafloor observatory	• Plate observing system	• Global ocean observing infrastructure	• Integrated carbon observation system	• Biodiversity and ecosystem research infra

Challenges

- Diverse standardises
  - Terminologies
  - Data models
  - Metadata
  - Service interfaces
  - …
- Semantic isolation
  - Between metadata
  - Between data content
  - Between services
  - …

Metadata standards	Acquisition	Curation	Processing	Access	Community
SensorML	Y	Y		Y	
NetCDF			Y	Y	
ISO19115	Y	Y		Y	Y
ISO19156	Y	Y			
CSR	Y	Y			
Dublin Core				Y	
CERIF	Y	Y		Y	
CSMD	Y	Y		Y	Y
INSPIRE	Y	Y		Y	Y

ENVRI project

- Cluster project for environmental ESFRIs
- Identify common operations and needs among ESFRIs
- Guide the development of ESFRIs
- Promote Interoperability
- Enable interdisciplinary scientists to access, process, study and correlate data from multiple domains for *system level* research.

Key problems and approaches

- Identify common requirements and operations from RIs → ODP approach
- Guide the RI development → Reference model approach
- Semantic gaps between RIs → ontological framework (OEILM)

Identify common requirements and operations

- Approach
  - Collect use cases and requirements from all ESFRIs
  - Analyse the use cases using the ODP approach
  - Define a minimal set which crosses most of the ESFRIs

Common Functions (Curation)

- Functions/operations in the Data Curation Sub-system

Functions/Embedded Services	ICOS	EPOS	EMSO	EISCAT-3D	LifeWatch	EURO-Argo
Data Quality Checking	Yes	Yes	Unknown	Yes	Not Applicable	Yes
Data Quality Verification	Yes	Unknown	Unknown	Unknown	Not Applicable	Yes
Data Identification	Yes	Yes	Yes	Unknown	Not Applicable	Unknown
Data Cataloguing	Unknown	Yes	Yes	Unknown	Not Applicable	Unknown
Data Product Generation	Yes	Yes	Yes	Yes	Not Applicable	Yes
Data Versioning	Yes	Unknown	Unknown	Unknown	Not Applicable	Unknown
Workflow Enactment	No	Yes	Unknown	Yes	Not Applicable	No
Data Preservation	Yes	Yes	Yes	Yes	Not Applicable	Yes
Data Replication	No	Yes	Unknown	Yes	Not Applicable	Yes
Data Replication Synchronisation	No	Unknown	No	Unknown	Not Applicable	Yes

Common Functions (Access)

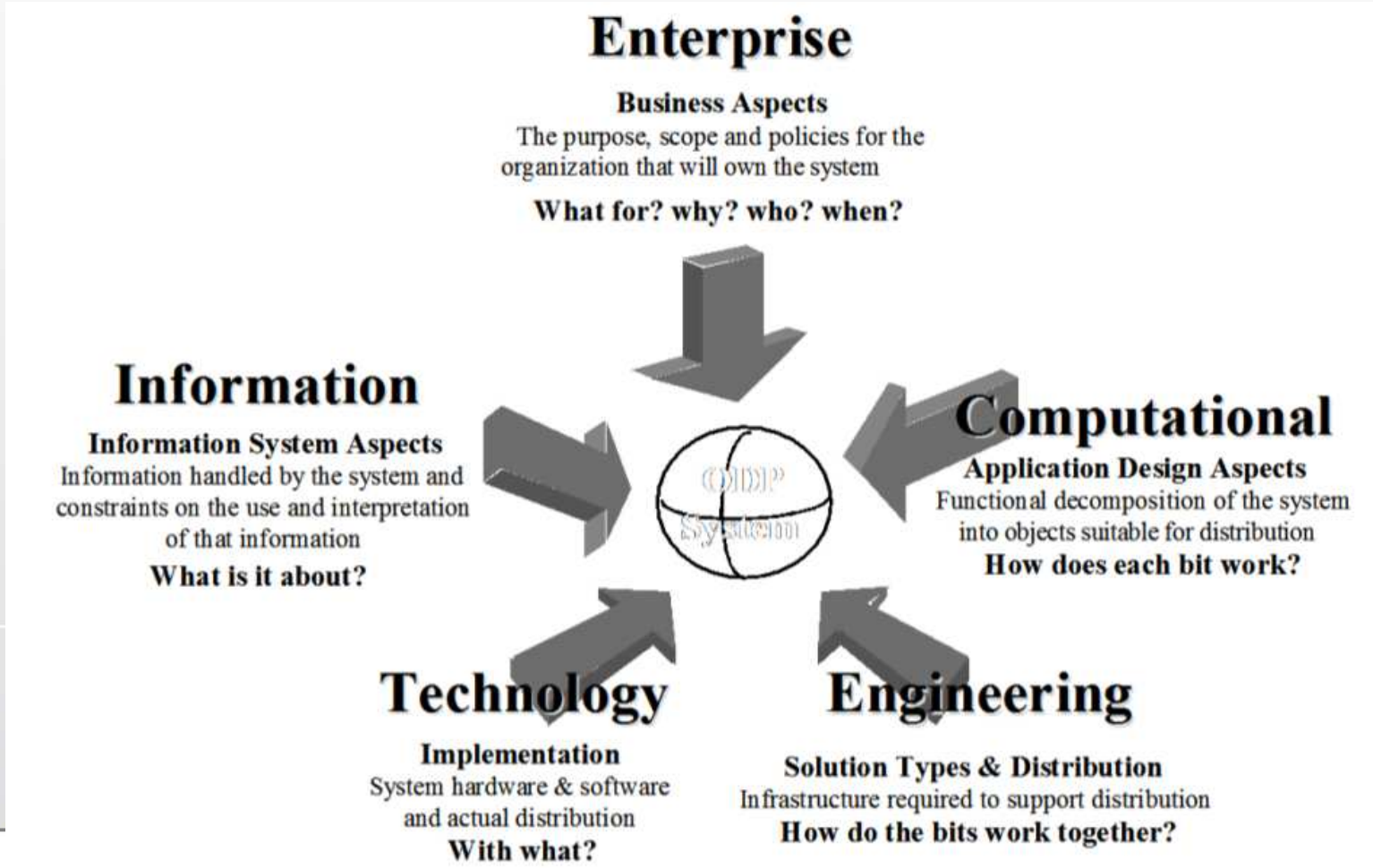
- **Functions/operations at Data Access Sub system**

Functions/Embedded Services	ICOS	EPOS	EMSO	EISCAT-3D	LifeWatch	Euro-Argo
Access Control	Unknown	Yes	Unknown	Yes	Unknown	Unknown
Data Conversion	Yes	Yes	Yes	Yes	Yes	Yes
Data Compression	No	No	No	No	Yes	No
Data Visualisation	Yes	Yes	Yes	Yes	Yes	Yes
Data Publication	Yes	Unknown	Yes	Unknown	Yes	Yes
Data Citation	No	Unknown	Yes	No	Unknown	No
(Resources/Data) Annotation	Yes	Yes	Yes	No	Yes	Yes
Metadata Harvesting	Unknown	Unknown	Yes	No	Unknown	No
Resource Registration	Unknown	Yes	Yes	No	Yes	No
Semantic Harmonisation	No	Yes	Yes	No	Yes	No
Data Discovery and Access	Yes	Yes	Yes	Yes	Yes	Unknown

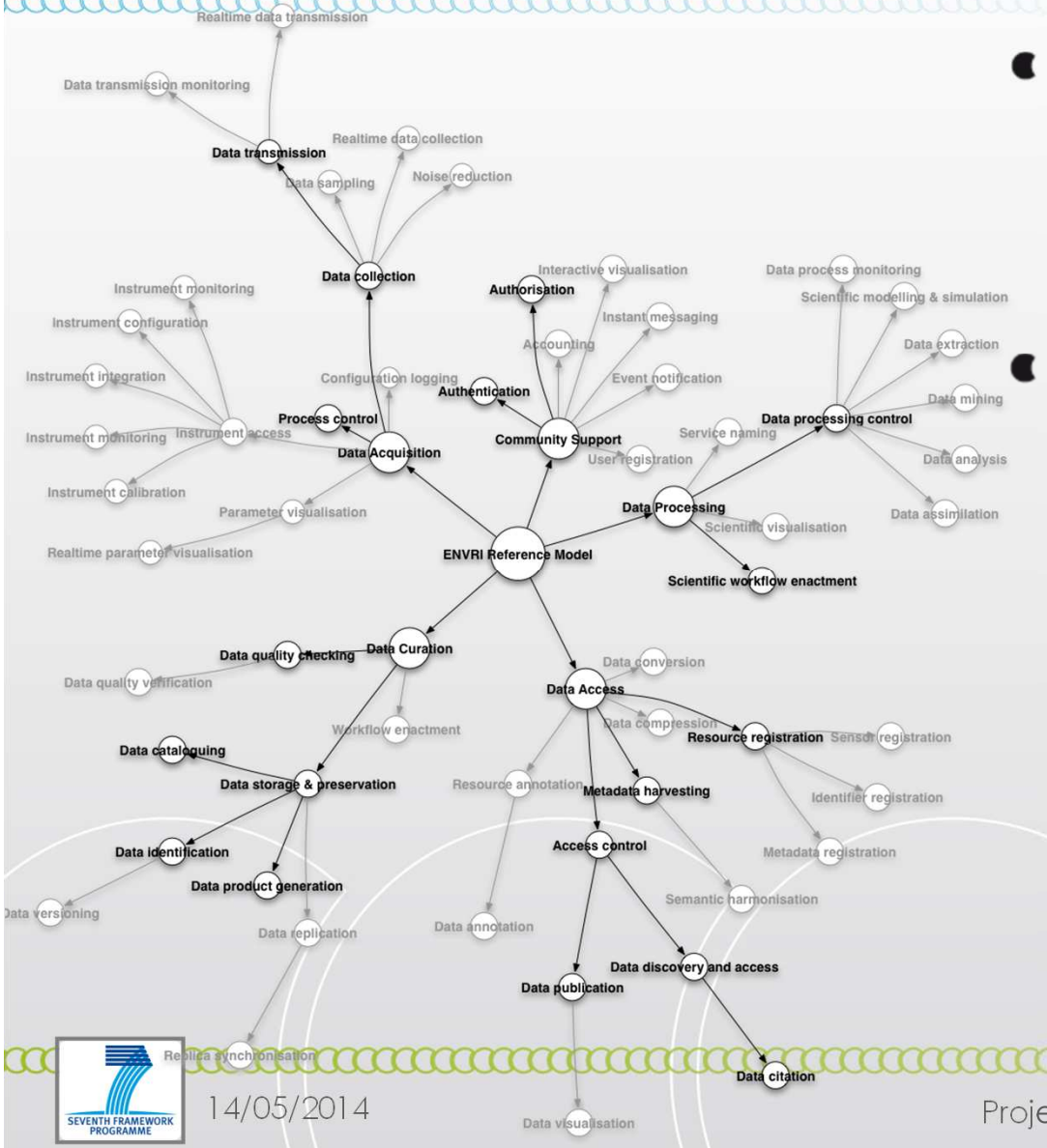
- **A full function list is on ENVRI wiki** http://envri.eu/group/envri/wiki/-/wiki/Main/Analyse%20Common%20Requirements%20for%20Data%20Processing

Analysis method

- **Open distributed processing (ODP): a multi viewpoint model for distributed systems.** (ISO/IEC 10746)



Identify a minimal set



- Analysis of common requirements of ESFRI ENV RIs, resulted in **a set of common functionalities**
- Identified **a *minimal model***
  - Focuses on **core interactions**
  - Represents the **most fundamental** functionalities
  - A **skeleton** which can be **extended**
  - Future development will be based on **community interests**

ENVRI RM: Science Viewpoint

- **Derive use scenarios from common requirements, identifying *communities*, *roles*, *behaviours***
- **Model defines:**
  - **5 common *Communities*** in according to 5-subsystem
    - **Data Acquisition:** who collects raw data
    - **Data Curation:** who manages, archives quality data
    - **Data Publication:** who assists publication, discovery & access
    - **Data Service Provision:** who provides services to derive knowledge
    - **Data Usage:** who makes use of data/services
  - **Community *roles & behaviours***

ENVRI RM: Information Viewpoint

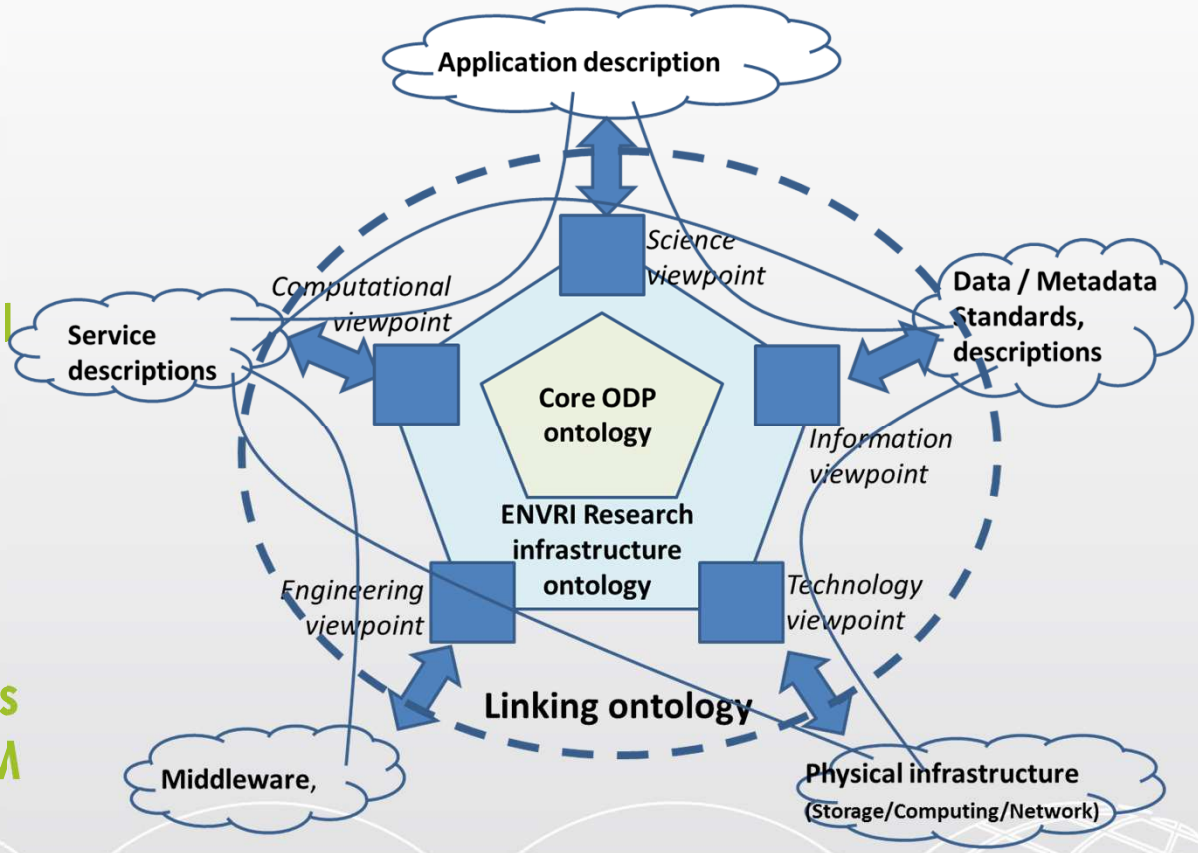
- **Data-oriented approach:**
  - Follow **data-lifecycle** in each subsystems
  - Identify **information objects**, **actions**, **state changes** when events/actions occur
- **Model defines:**
  - A set of **information objects** handled by a subsystem
  - A set of **action types** that cause the states changes
  - A set of **constraints** on those objects
  - **Dynamic schema** -- how information objects evolve as the system operates
  - **Static schema** -- allowable state changes

ENVRI RM: Computational VP

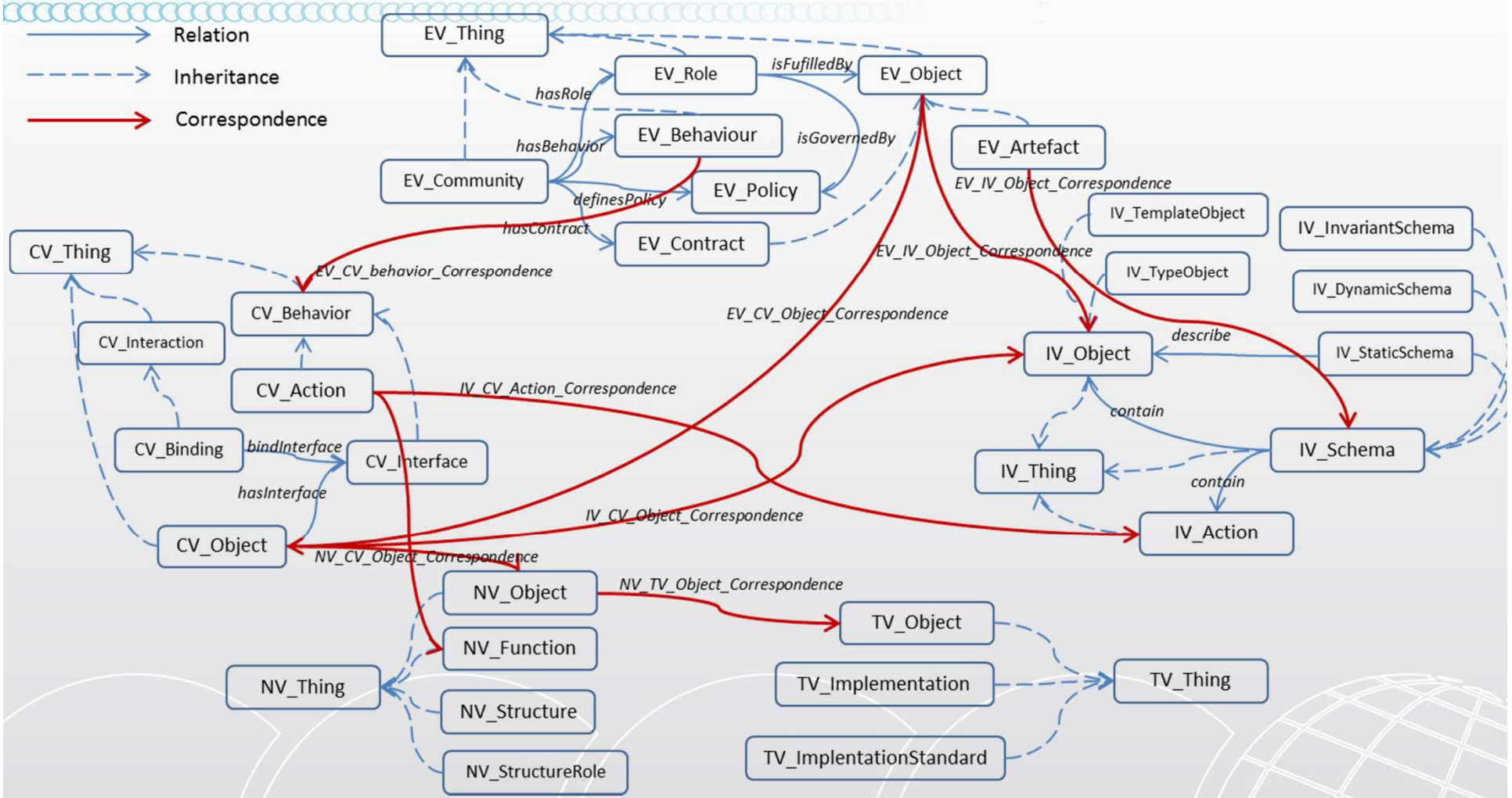
- **Service-oriented, Brokered approach**
  - **Core functionality is encapsulated in** a set of **service** objects
  - **Access** to such object **via brokers** which provides an **interoperability layer** between heterogeneous components
- **Model defines**
  - A set of **computational objects**
    - Each **encapsulates** specific functionalities
    - Each provides a set of **interfaces** to invoke functions
  - A set of **binding objects** to coordinate multi-party interactions

OEILM: Ontological framework

- ODP ontology
  - Vocabulary in ODP
  - Five viewpoints
- ENVRI RM ontology
  - Vocabulary in ENVRI RM
  - Five viewpoints extended from ODP
- Linking ontology
  - Concepts/properties extend the ENVRI RM from five viewpoints

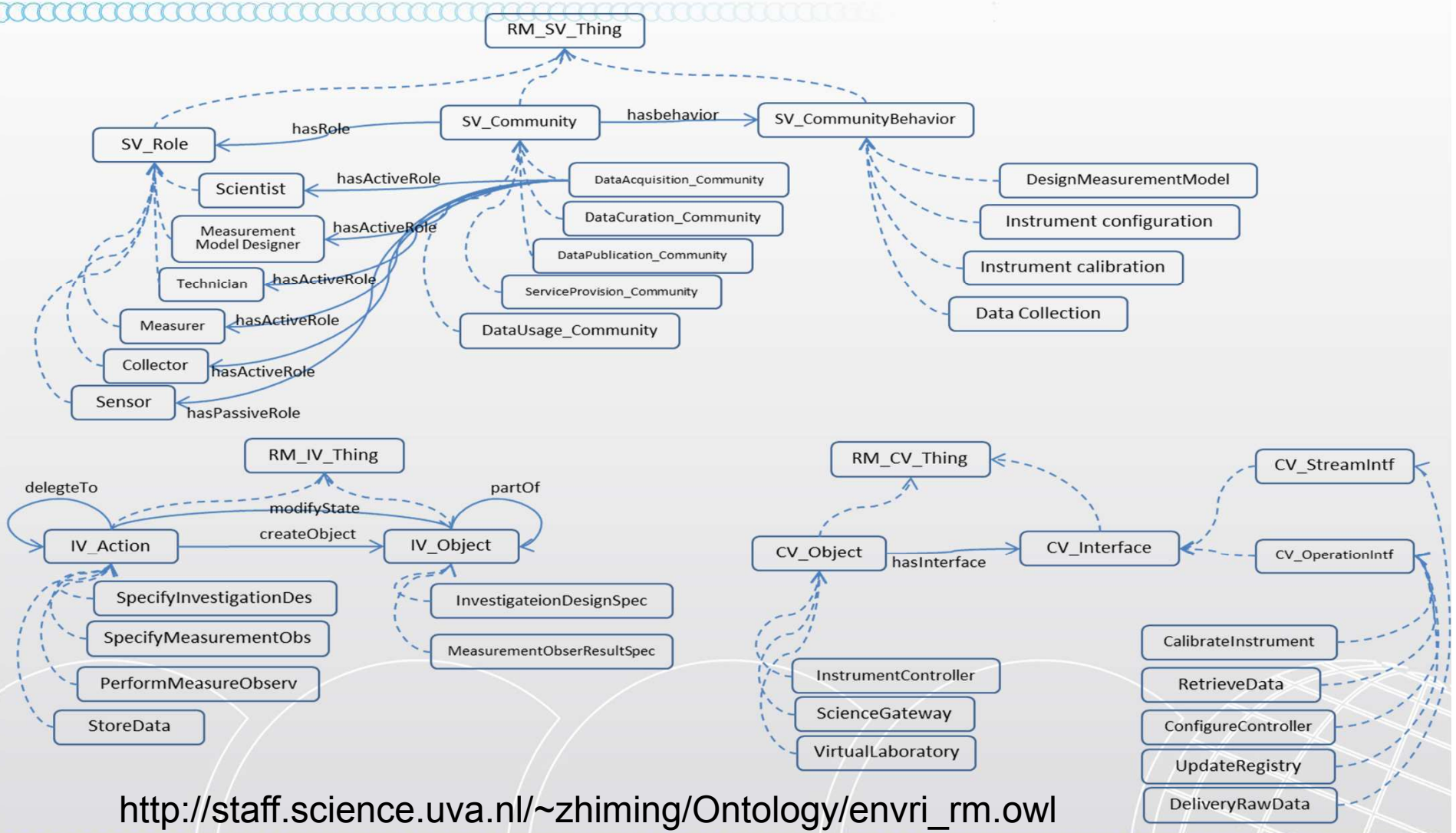


ODP core ontology (part)



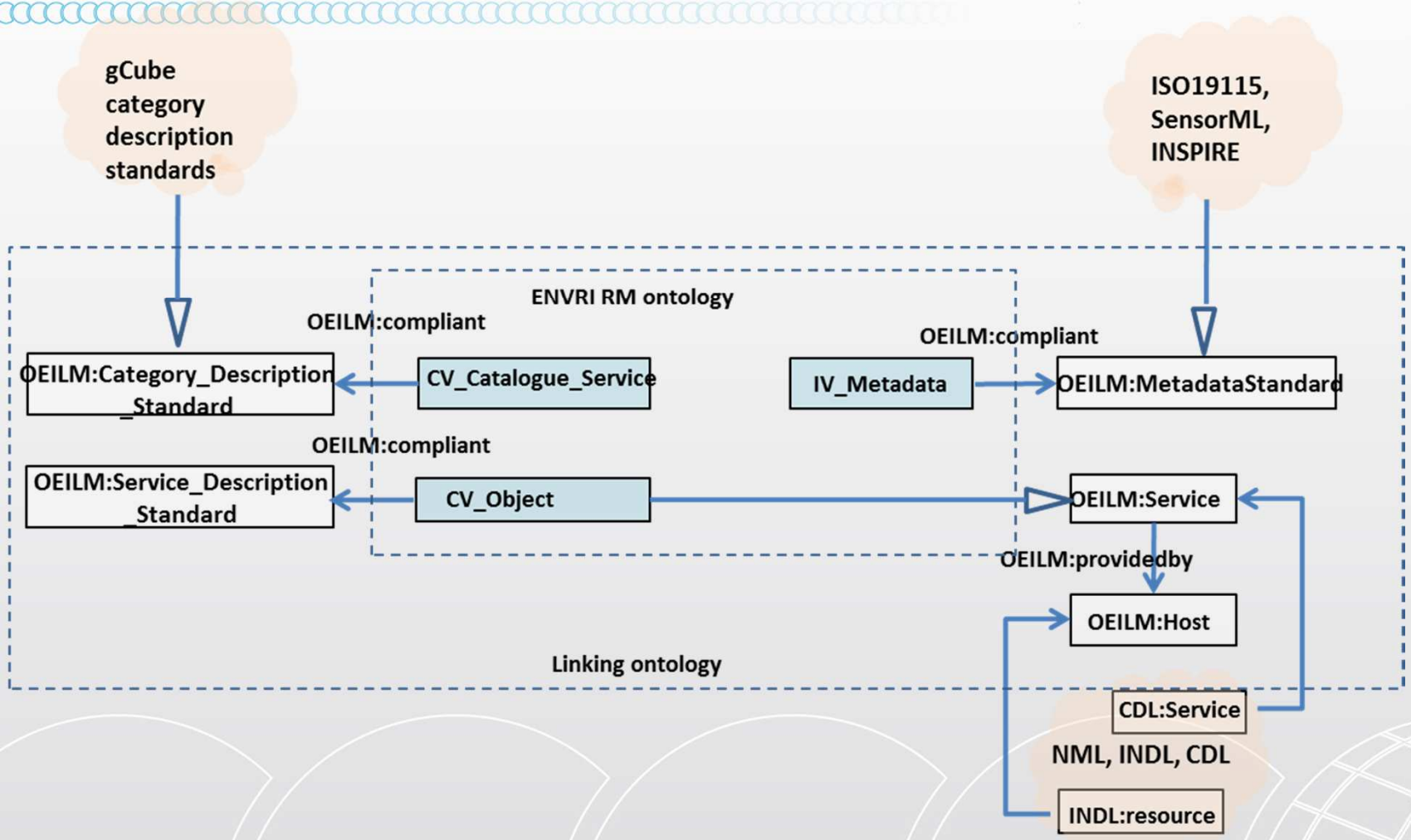
http://staff.science.uva.nl/~zhiming/Ontology/odp.owl

ENVRI RM ontology (part)

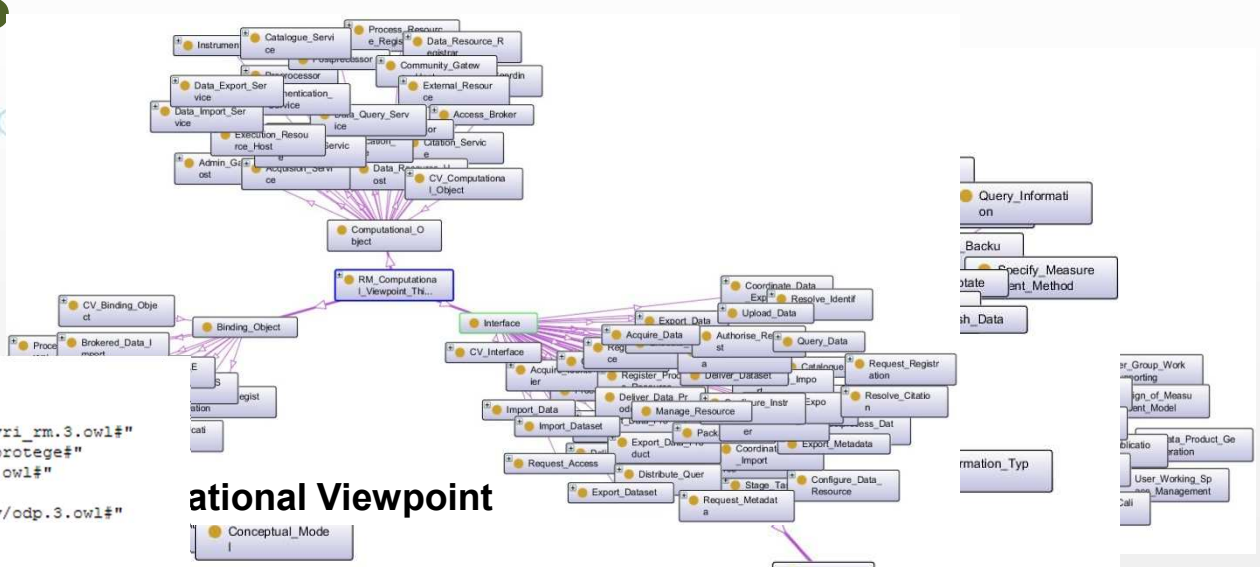


http://staff.science.uva.nl/~zhiming/Ontology/envri_rm.owl

Linking ontology (part)

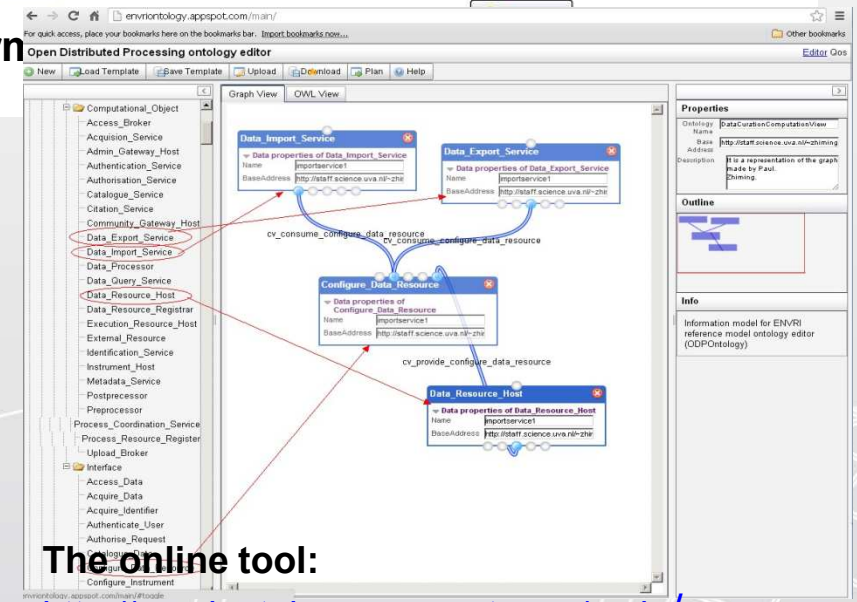


Tools



ational Viewpoint

Inform




```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns="http://staff.science.uva.nl/~zhiming/ontology/envri_rm.3.owl#"
    xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
    xmlns:xsp="http://www.owl-ontologies.com/2005/08/07/xsp.owl#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:odp="http://staff.science.uva.nl/~zhiming/ontology/odp.3.owl#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:swrl="http://www.w3.org/2003/11/swrl#"
    xmlns:swrlb="http://www.w3.org/2003/11/swrlb#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xml:base="http://staff.science.uva.nl/~zhiming/ontology/envri_rm.3.owl">
  <owl:Ontology rdf:about="">
    <owl:versionInfo xml:lang="en">Version 1</owl:versionInfo>
    <Contributer rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
    >Creater:
Zhiming Zhao, Z.Zhao@uva.nl, University of Amsterdam</Contributer>
    <rdfs:comment xml:lang="en">&lt;p style="margin-top: 0"&gt;
      Log
    &lt;/p&gt;
    &lt;p style="margin-top: 0"&gt;
      date: 26-march-2013
    &lt;/p&gt;
    &lt;p style="margin-top: 0"&gt;
      Computational view:
    &lt;/p&gt;
    &lt;p style="margin-top: 0"&gt;
      - computational ojbects have been added
    &lt;/p&gt;
    &lt;p style="margin-top: 0"&gt;
      - interface have been added
    &lt;/p&gt;
    &lt;p style="margin-top: 0"&gt;
      - properties to the interface is modelled as sub property of interface
    &lt;/p&gt;
    &lt;p style="margin-top: 0"&gt;
      - in the odp ontology, has interface will have two sub properties:
                                    _interface
    &lt;/p&gt;
```

RM Owl version:
http://staff.science.uva.nl/~zhiming/Ontology/http://envriontology.appspot.com/main/.

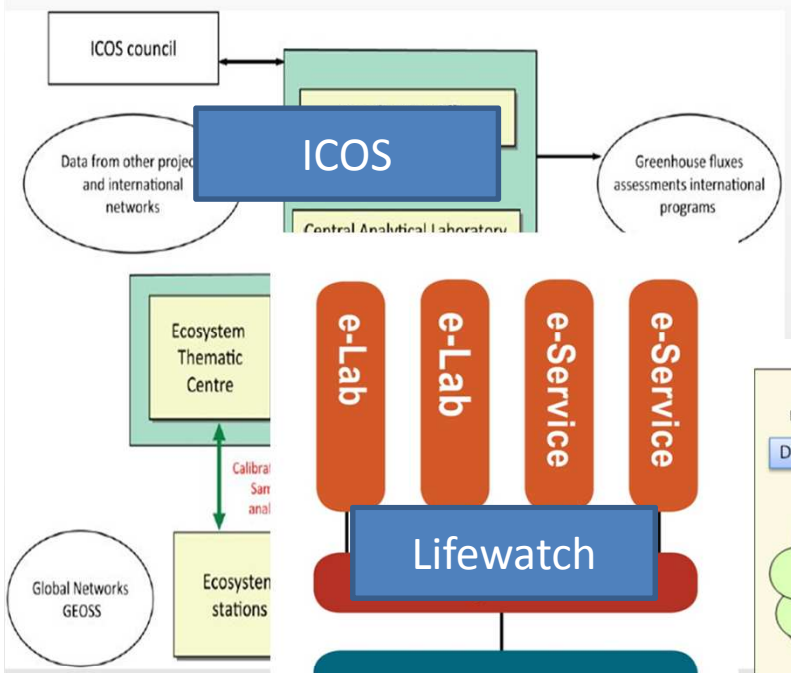
The online tool:
http://envriontology.appspot.com/main/.

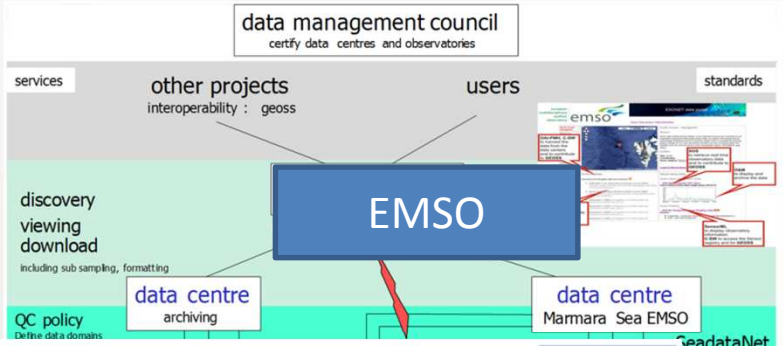
Use cases of OEILM

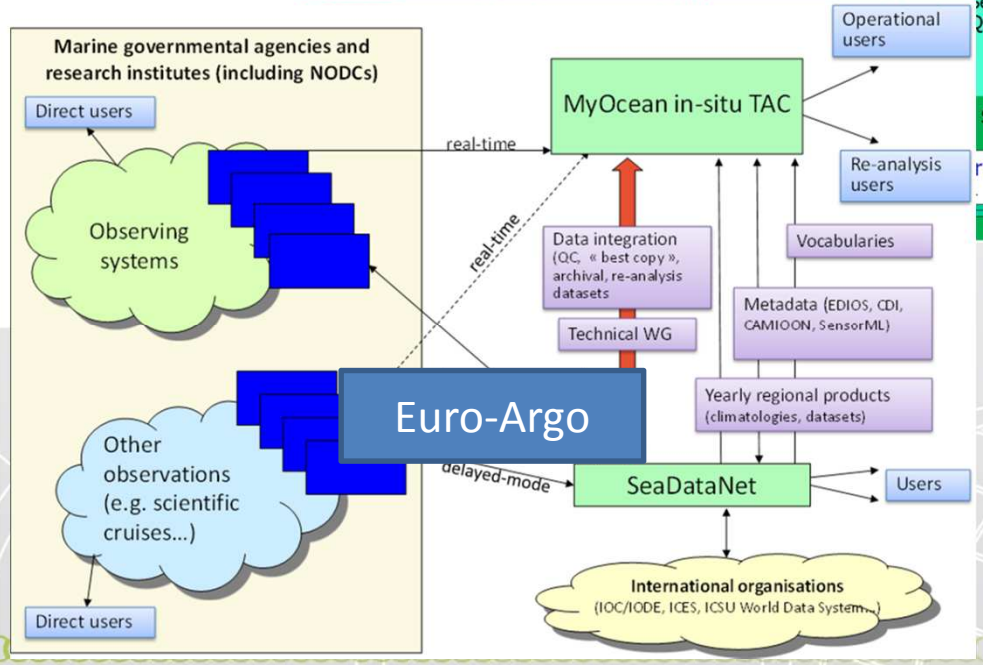
- **ESFRI information sharing**
- **Resource discovery and workflow planning**

Use case 1:Sharing design documents between ESFRIs

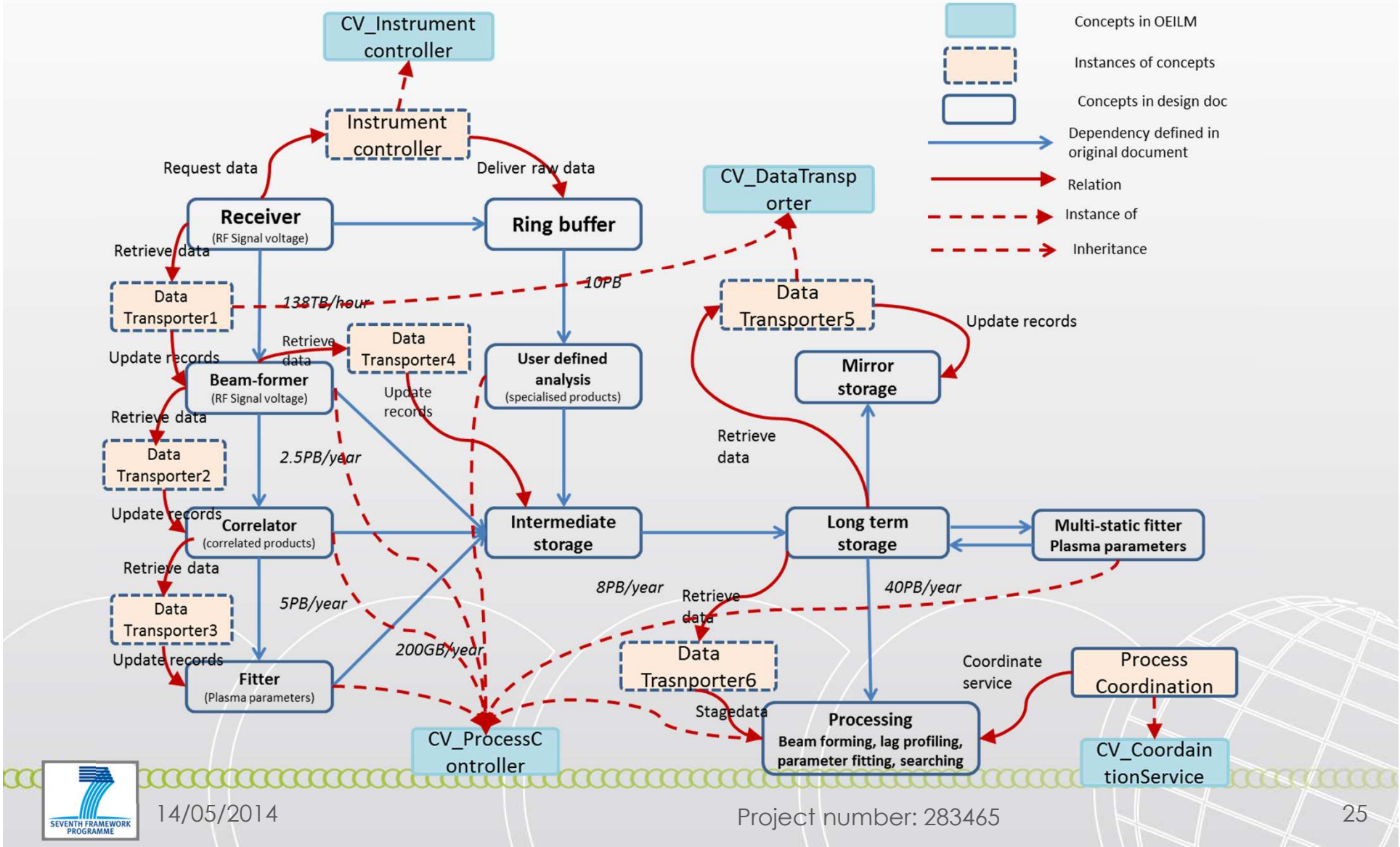
- **Different terminology makes sharing difficult**



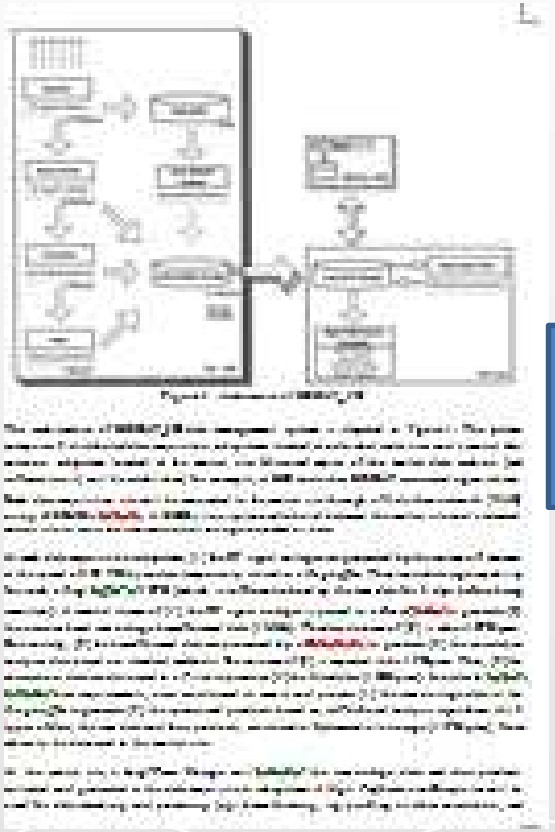




Annotating ESFRI documents using OEILM

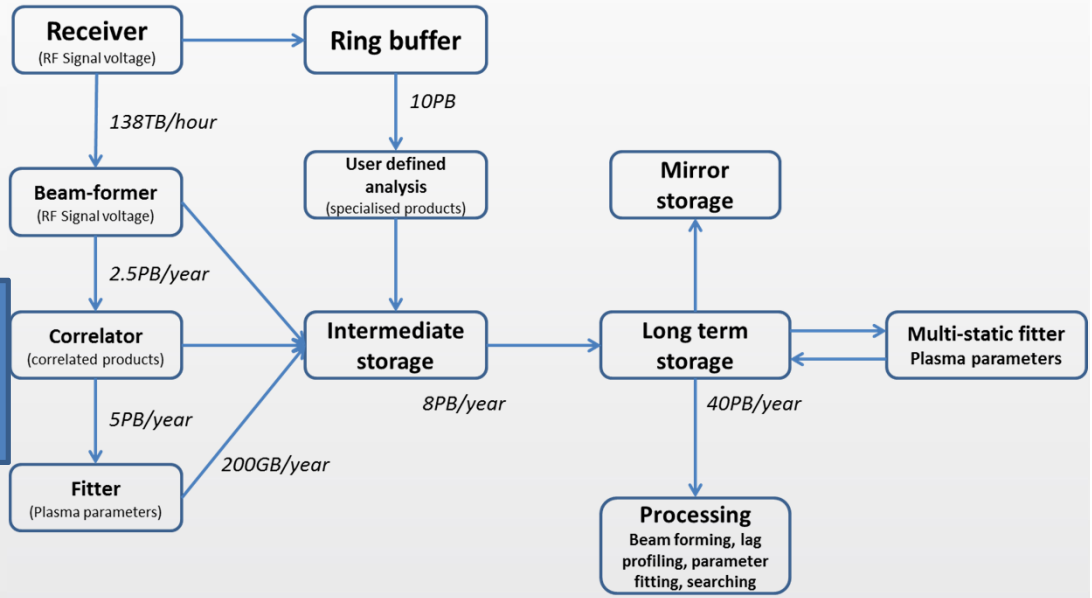


Step 1: design doc formalization



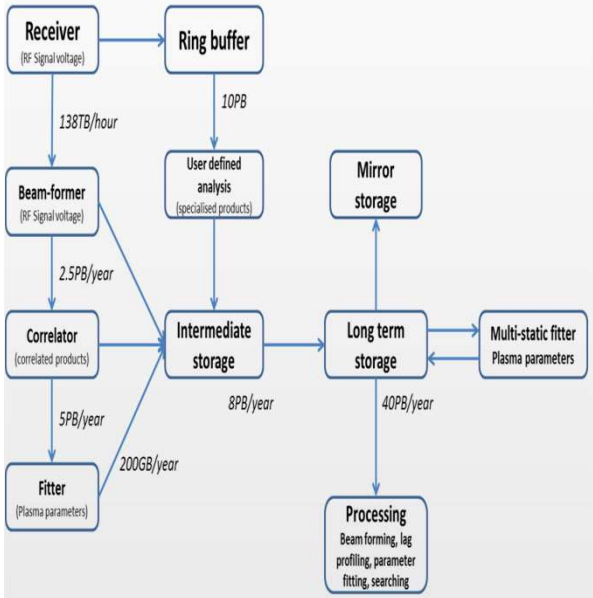
Provided by RI designers,
http://envri.eu/eiscat_3d-study-case

Step 1 ➔

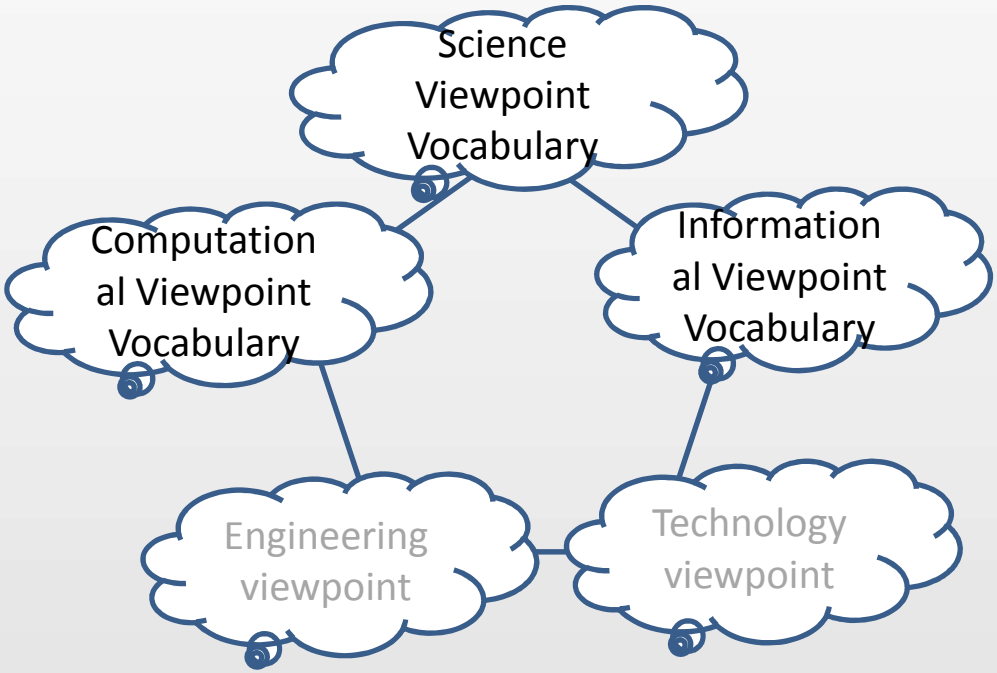


Identify the processes/data from the design documents

Step 2: find proper OEILM classes/properties (RM layer)



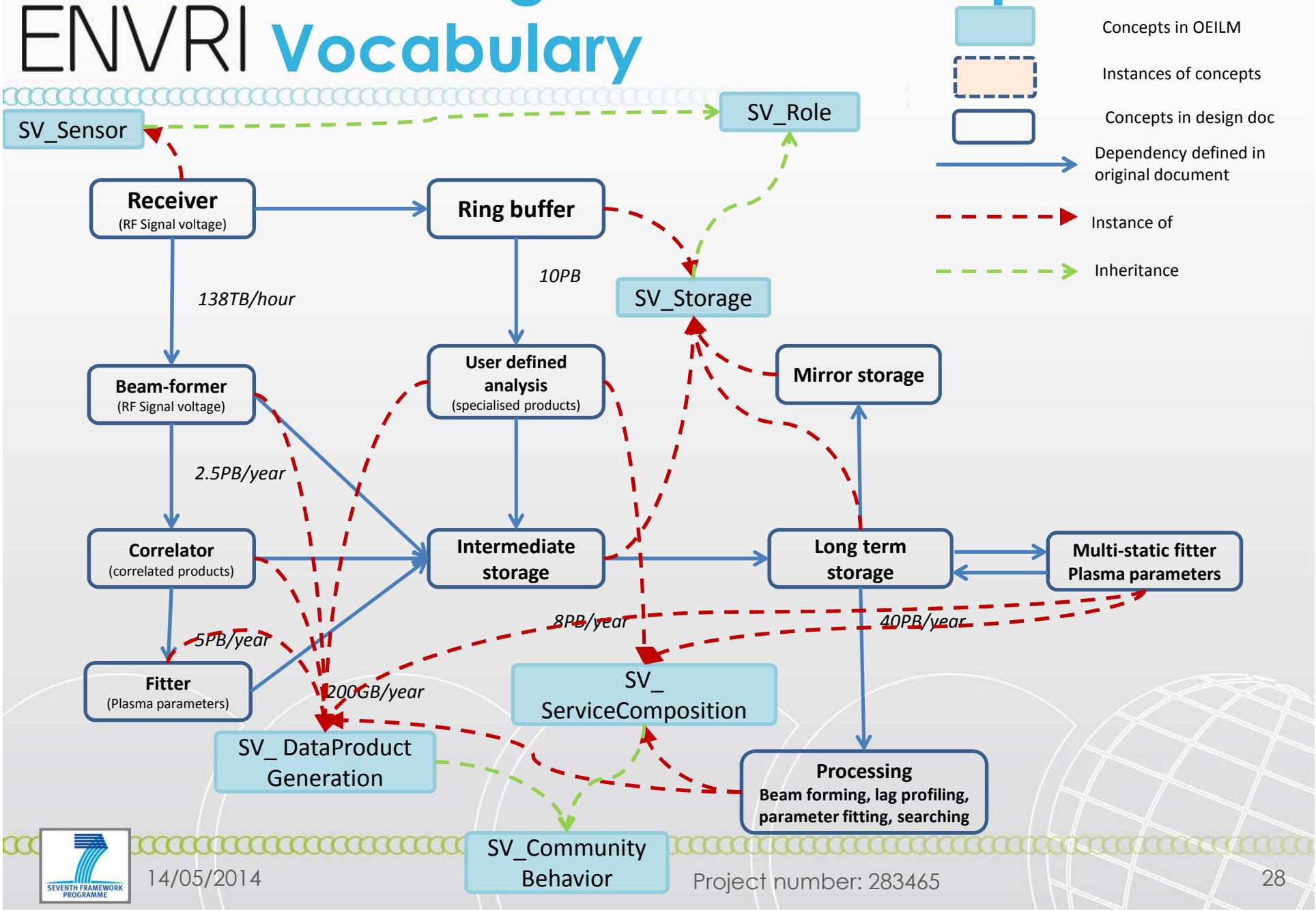
Step 2 ➔



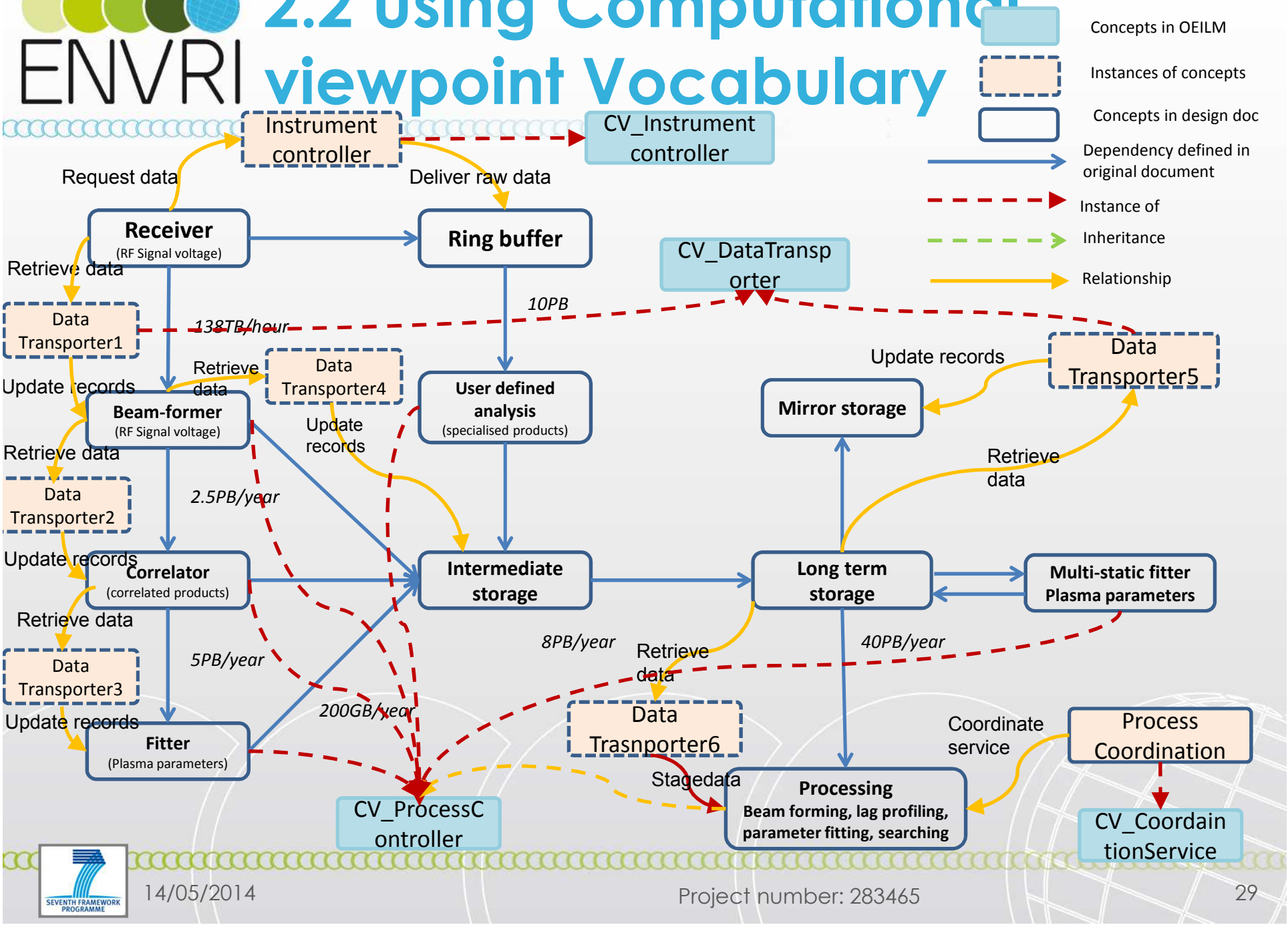
Map the concepts/data to the ENVRI RM concepts

1) **Find the instances/classes**
2) **Identify the missing ones between instances**

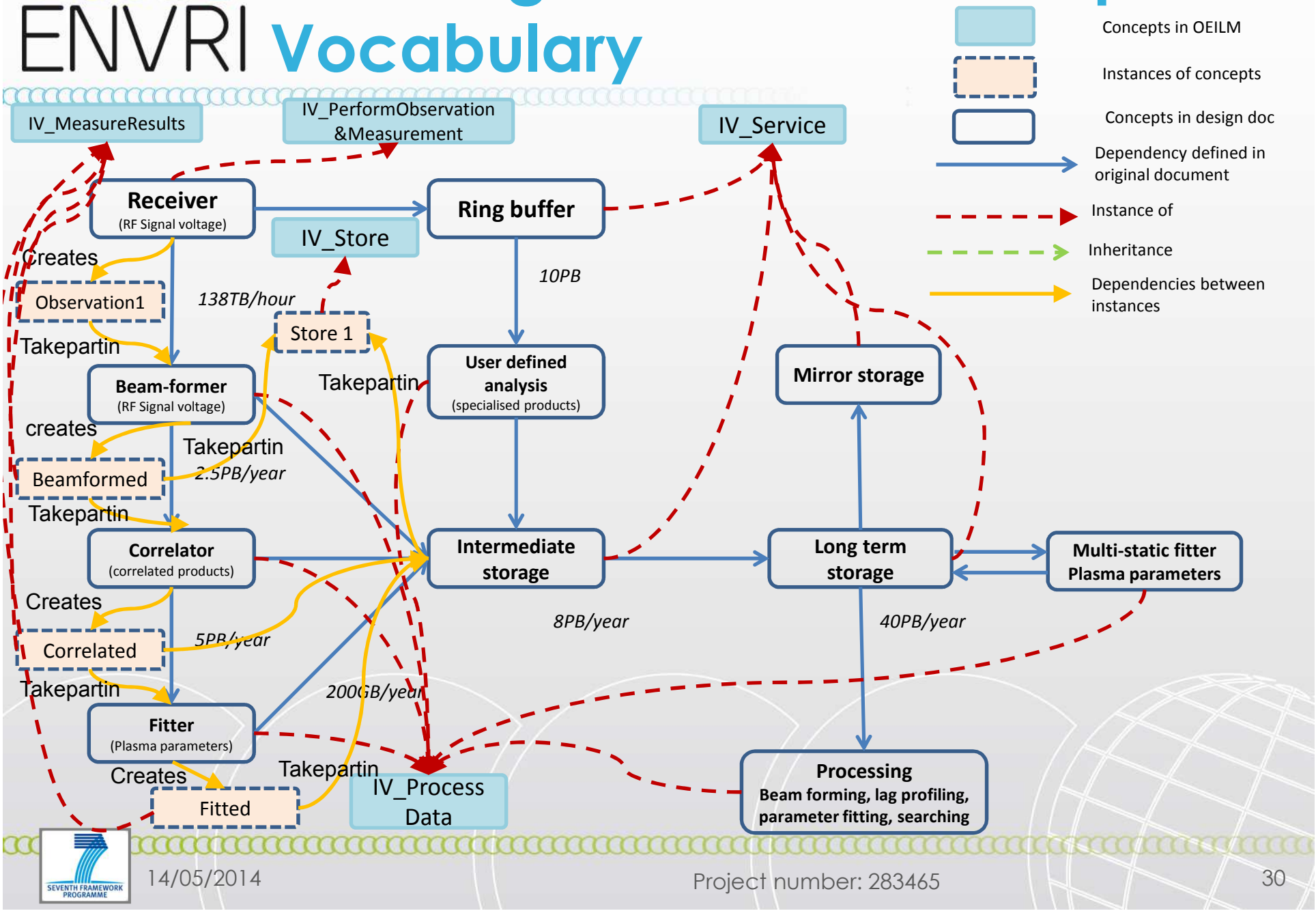
2.1 Using Science viewpoint Vocabulary



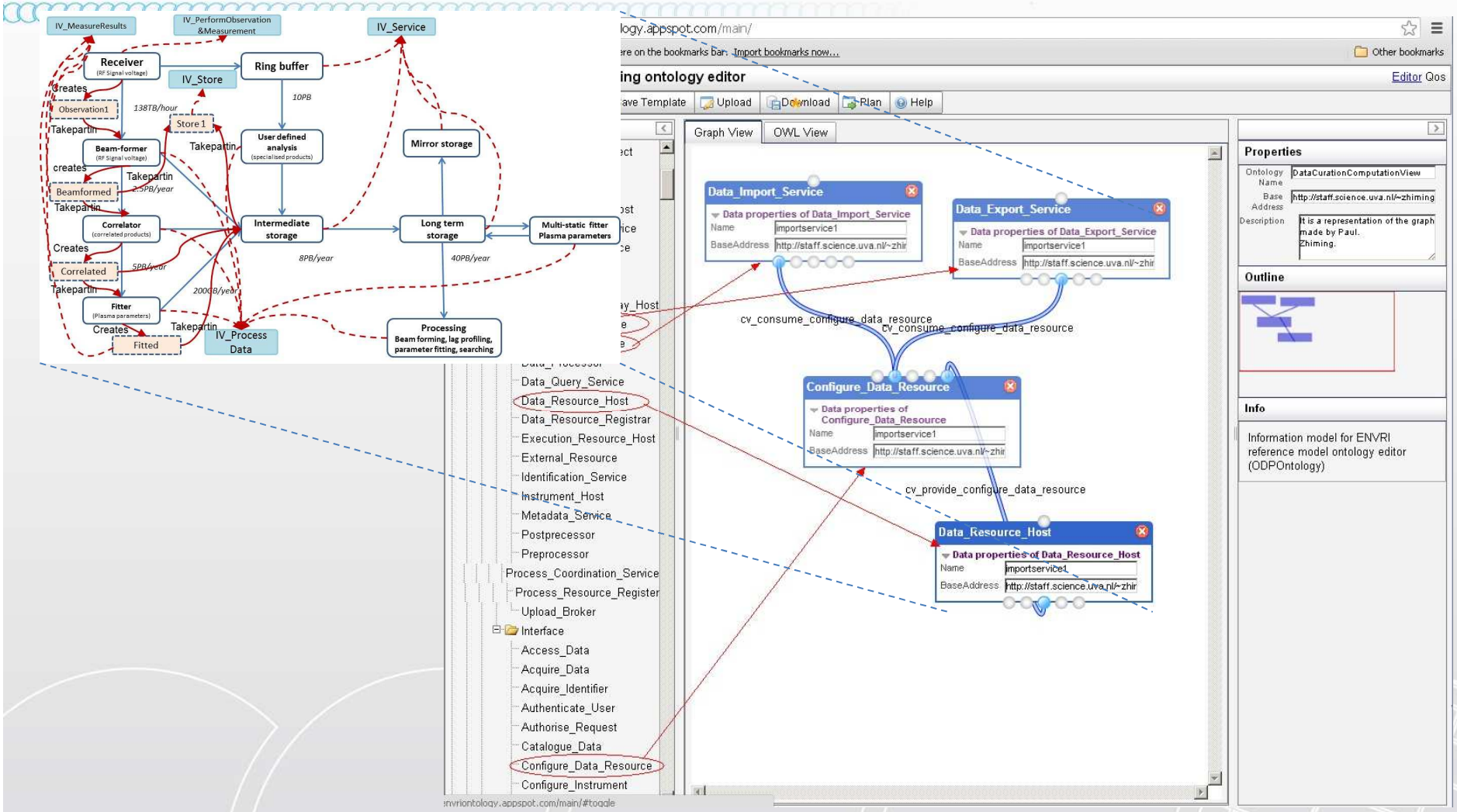
2.2 Using Computational viewpoint Vocabulary



2.3 Using Information viewpoint Vocabulary



OEILM tools for annotation/description

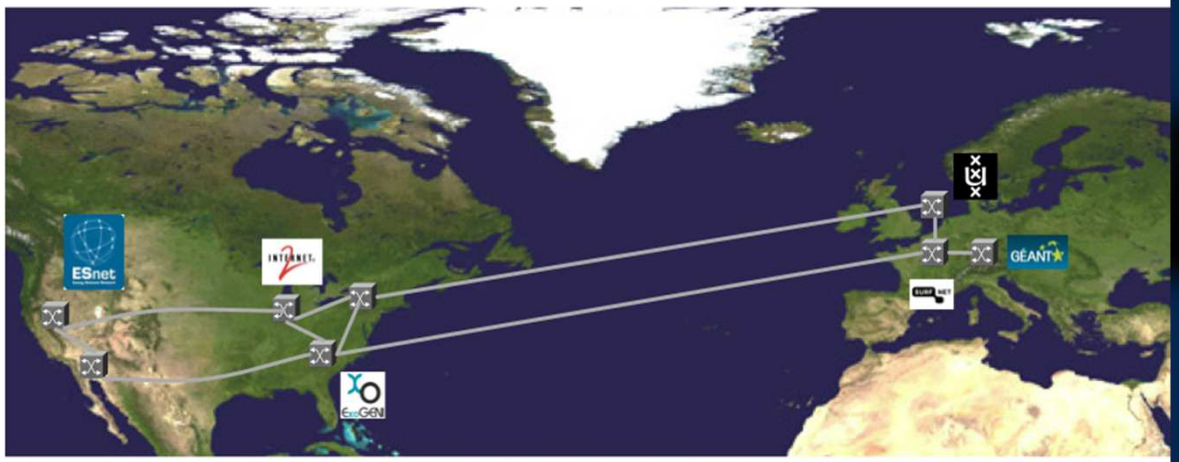


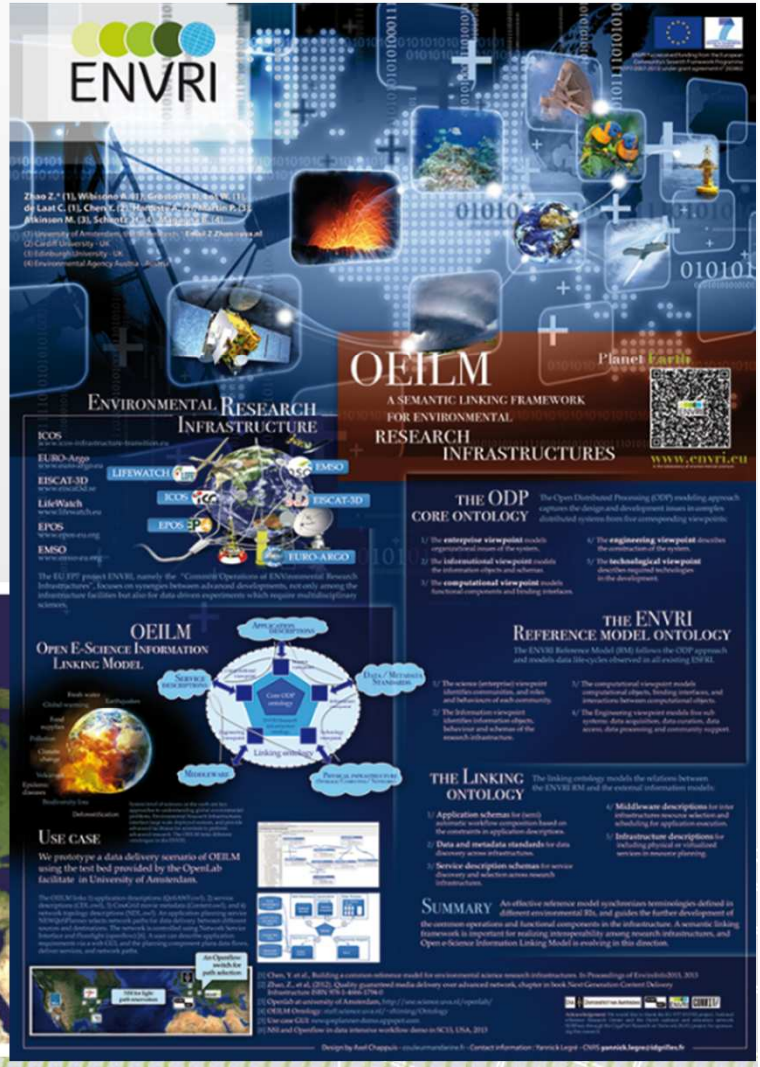
Step 3: Match-making

- From the annotation, we can identify the main functional components:
  - Science Viewpoint
    - SV_Role (SV_Sensor, SV_Storage)
    - SV_CommunityBehavior(SV_ProductGeneration, SV_ServiceComposition)
  - Computational Viewpoint
    - CV_InstrumentController, CV_DataTransporter, CV_ServiceCoordinator, CV_ProcessController
  - Information Viewpoint
    - IV_MeasurementResults, IV_Store, IV_Service, IV_ProcessData
- Functional requirements:
  - Acquisition:
    - SV_Sensor- CV_InstrumentController- IV_MeasurementResults
  - Curation
    - SV_Stoage, CV_Transporter, ***IV_Annotation, IV_PersistentIdentifier, IV_Metadata***
  - Access
    - SV_Storage, CV_Transporter, ***IV_Metadata, IV_Identifier,***
  - Processing
    - SV_ProductionGeneration, SV_ServiceComposition, CV_Staging, CV_ServiceCoordinator, CV_ProcessController, IV_ProcessData, IV_Service,
  - Community
- EGI infrastructure
  - Curation (annotation, replication, catalogue)
  - Processing (computing power, parameter sweeping tool, vis, etc.)
  - Access (metadata, catalogue, AAA, storage, transfer)

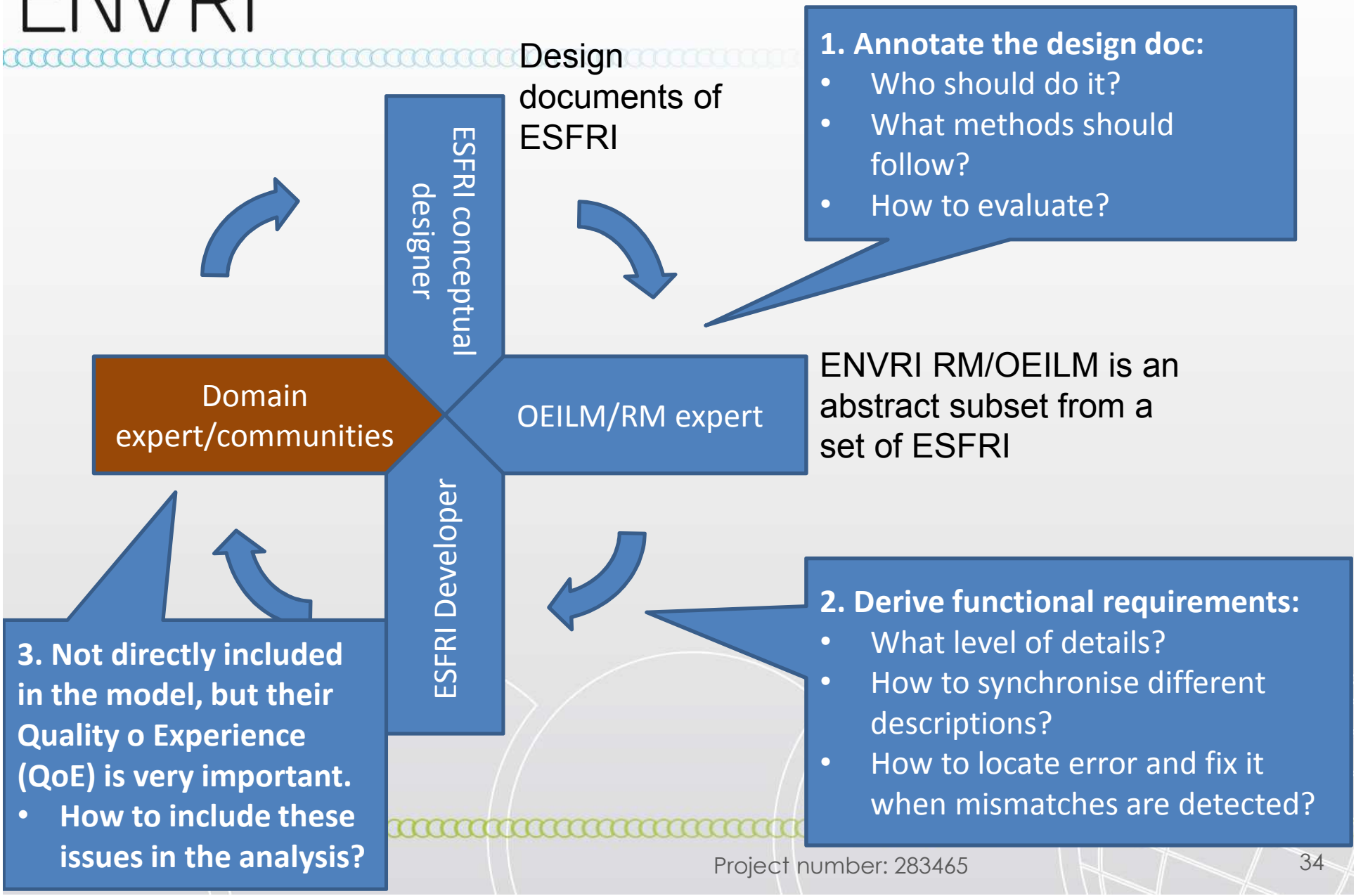
Use case 2: application planning using linking ontology

- Application QoS ←→ Services /Data ←→ Devices ←→ Network paths
- Live demo in SC 13





Discussion


Design documents of ESFRI
ESFRI conceptual designer
Domain expert/communities
OEILM/RM expert
ESFRI Developer
ENVRI RM/OEILM is an abstract subset from a set of ESFRI
1. Annotate the design doc:
• Who should do it?
• What methods should follow?
• How to evaluate?
2. Derive functional requirements:
• What level of details?
• How to synchronise different descriptions?
• How to locate error and fix it when mismatches are detected?
3. Not directly included in the model, but their Quality o Experience (QoE) is very important.
• How to include these issues in the analysis?

Summary

- The ENRI project will finish by the end of 2014
- The first version of OEILM has been made available for ENVRI ESFRI
- Ongoing work
  - Use cases with each specific ESFRI
  - Transfer knowledge to ESFRI
  - Collect feedback and refine OEILM for its version 2.
- Dissemination
  - RDA/EUDAT
- Exploitation
  - ENVRI 2 proposal

References

- ENVRI: www.envri.eu
- OEILM: http://staff.science.uva.nl/¬zhiming/Ontology

- Zhao Z., Gross, P., Wouter. L., Chen Y., Hardisty, A., Martine, P., Magana, B., Schentz, H., (2013) *OEILM: a semantic linking framework for environmental research infrastructures*, Poster at IEEE e-Science 2013.
- Jiang, W., Zhao, Z., Grosso, P., de Laat, C., (2013) *Dynamic workflow planning on programmable infrastructure*, IEEE Int'l Conf. on Network Architecture Storage, 2013.
- Chen, Y., Martin, P., Schentz, H., Magagna, B., Zhao, Z., Hardisty, A., Preece, A., Atkinson, M., Huber, R., and Legre. R., *Building a common reference model for environmental science research infrastructures*. In Proceedings of EnviroInfo2013, 2013.
- Zhao, Z., van der Ham, J., Taal, A., Koning, R., Dumitru, C., Wibisono, A., Grosso, P., de Laat., C. (2012). *Planning data intensive workflows on inter-domain resources using the Network Service Interface (NSI)*, In the 7th Workshop on Workflows in Support of Large-Scale Science. in the context of Supercomputing 2012, Salt Lake City.
- Zhao, Z., Grosso, P. & Laat, C. de (2012). *OEIRM: An Open Distributed Processing based Interoperability Reference Model for e-Science*, Cloud&Grid interoperability workshop, Gwangju, Korean.
- Zhao, Z., Dumitru, C., Grosso, P. & Laat, C. de (2012). *Network resource control for data intensive applications in heterogeneous infrastructures*. the International Workshop on High Performance Data Intensive Computing, in 26th IEEE International Parallel and Distributed Processing Symposium, Shanghai.